

## Article

# Comparison of Masking Algorithms for Sentinel-2 Imagery

Viktoria Zekoll <sup>1,\*</sup>, Magdalena Main-Knorn <sup>2</sup>, Jerome Louis <sup>3</sup> , David Frantz <sup>4</sup> , Rudolf Richter <sup>1</sup>   
and Bringfried Pflug <sup>2</sup>

<sup>1</sup> DLR, German Aerospace Center, D-82234 Wessling, Germany; Rudolf.Richter@dlr.de

<sup>2</sup> DLR, German Aerospace Center, D-12489 Berlin, Germany; magdalena.main-knorn@dlr.de (M.M.-K.); Bringfried.Pflug@dlr.de (B.P.)

<sup>3</sup> Telespazio France, 31023 Toulouse, France; jerome.louis@telespazio.com

<sup>4</sup> Earth Observation Lab, Geography Department, Humboldt-Universität zu Berlin, D-10099 Berlin, Germany; david.frantz@geo.hu-berlin.de

\* Correspondence: viktoria.zekoll@dlr.de

**Abstract:** Masking of clouds, cloud shadow, water and snow/ice in optical satellite imagery is an important step in automated processing chains. We compare the performance of the masking provided by Fmask (“Function of mask” implemented in FORCE), ATCOR (“Atmospheric Correction”) and Sen2Cor (“Sentinel-2 Correction”) on a set of 20 Sentinel-2 scenes distributed over the globe covering a wide variety of environments and climates. All three methods use rules based on physical properties (Top of Atmosphere Reflectance, TOA) to separate clear pixels from potential cloud pixels, but they use different rules and class-specific thresholds. The methods can yield different results because of different definitions of the dilation buffer size for the classes cloud, cloud shadow and snow. Classification results are compared to the assessment of an expert human interpreter using at least 50 polygons per class randomly selected for each image. The class assignment of the human interpreter is considered as reference or “truth”. The interpreter carefully assigned a class label based on the visual assessment of the true color and infrared false color images and additionally on the bottom of atmosphere (BOA) reflectance spectra. The most important part of the comparison is done for the difference area of the three classifications considered. This is the part of the classification images where the results of Fmask, ATCOR and Sen2Cor disagree. Results on difference area have the advantage to show more clearly the strengths and weaknesses of a classification than results on the complete image. The overall accuracy of Fmask, ATCOR, and Sen2Cor for difference areas of the selected scenes is 45%, 56%, and 62%, respectively. User and producer accuracies are strongly class- and scene-dependent, typically varying between 30% and 90%. Comparison of the difference area is complemented by looking for the results in the area where all three classifications give the same result. Overall accuracy for that “same area” is 97% resulting in the complete classification in overall accuracy of 89%, 91% and 92% for Fmask, ATCOR and Sen2Cor respectively.

**Keywords:** Sentinel-2; masking; Fmask; ATCOR; Sen2Cor



**Citation:** Zekoll, V.; Main-Knorn, M.; Louis, J.; Frantz, D.; Richter, R.; Pflug, B. Comparison of Masking Algorithms for Sentinel-2 Imagery. *Remote Sens.* **2021**, *13*, 137. <https://doi.org/10.3390/rs13010137>

Received: 1 December 2020

Accepted: 27 December 2020

Published: 4 January 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The Sentinel-2 mission consists of two polar-orbiting satellites, Sentinel-2A and Sentinel-2B, providing a five day revisit time at the equator. The swath width of a Sentinel-2 scene is 290 km and data is acquired in 13 bands with spatial resolutions of 10 m, 20 m, and 60 m [1] (see Table 1). Sentinel-2 images are open access data, offer high quality radiometric measurements and include a dedicated cirrus detection band. The free data access, frequent coverage of territories, wide swath and many spectral bands are reasons for the wide-spread use of this kind of data in many applications. Satellite imagery is frequently contaminated by low and medium altitude water clouds as well as by high-altitude cirrus clouds in the upper troposphere and in the stratosphere. Many operations require clear sky pixels as input, such as agriculture related products [2,3], the retrieval of surface reflectance within atmospheric correction [4,5] and the coregistration with other images [6,7].

**Table 1.** Sentinel-2 spectral bands and spatial resolution.

Sentinel-2 Bands	Resolution (m)
band 1 (0.433–0.453)	60
band 2 (0.458–0.523)	10
band 3 (0.543–0.578)	10
band 4 (0.650–0.680)	10
band 5 (0.698–0.713)	20
band 6 (0.733–0.748)	20
band 7 (0.765–0.785)	20
band 8 (0.785–0.900)	10
band 8a (0.855–0.875)	20
band 9 (0.930–0.950)	60
band 10 (1.365–1.385)	60
band 11 (1.565–1.655)	20
band 12 (2.100–2.280)	20

Atmospheric correction and land cover classification depend on an accurate cloud map [8–10]. In addition, maps of water and snow/ice are also indispensable in many applications, i.e., mapping of glaciers [11] and water bodies [12].

Cloud screening is applied to the data in order to retrieve accurate atmospheric and surface parameters as input for further processing steps, either the *Atmospheric Correction* (AC) itself or higher-level processing such as compositing, time-series analysis or estimation of biogeophysical parameters.

However, a fully automatic detection of these classes is not an easy task, due to the high reflectance variability of earth surfaces. For instance, bright desert surfaces or urban structures can be misclassified as one or the other or as cloud and shadow surfaces as water. A class assignment for mixed pixels (e.g., semitransparent cloud over snow) can be problematic because they do not have a spectral signature, which clearly belongs to a class. These together will decrease the classification accuracy and shows the need for a performance assessment of classification algorithms.

The *Cloud Masking Intercomparison Exercise* (CMIX) [13] was a recent state-of-the art intercomparison of a set of cloud detection algorithms for Sentinel-2 and Landsat-8 representative for sensors in the 10–30 m range. However, CMIX was limited to differentiate only cloudy and cloudless pixels. Reference [14] is limited to valid and invalid pixels too. Valid pixels in reference [14] are cloudless pixels like land, water and snow and invalid are clouds and cloud shadows. Cloud masks from the *MACCS ATCOR Joint Atmospheric Correction* (MAJA) algorithm using multi-temporal information are compared with monotemporal classification by Sen2Cor and Fmask [15]. The comparison in reference [14] is done twice: Once for cloud masks of all three processors dilated around clouds and second for all processors with nondilated cloud masks. This means that there is no comparison on original processor outputs. Overall accuracies for all three algorithms are nearby at 90–93% in case of nondilated cloud mask. Monotemporal Fmask gave equivalent classification performance as multitemporal MAJA for dilated masks and Sen2Cor was on average 6% worse on these. However, dilation of Sen2Cor cloud mask is not recommended with the used processor version because it is a known issue that it misclassifies many bright objects as clouds in urban area, which leads to commission of clouds and even more if dilation is applied. On the contrary, original masking outputs are evaluated in this paper and not only for valid and invalid pixels, but in more detail for six consolidated classes given below. This gives more insight into strengths and weaknesses of the masking algorithms.

As opposed to radiometric validation, the validation of masking is limited due to the lack of suitable reference datasets. Imaged-based reference data are required, which can only be generated through image interpretation or semiautomated methods as done in [14]. CMIX is based on four classification reference data bases for Sentinel-2 data. These existing

reference data are either not publicly available or do not fulfill the requirements for this study, e.g., 20 m resolution and a distinction of all defined classes.

In this study we evaluate the performance of three widely used monotemporal masking codes on Sentinel-2 imagery.

Our first masking code is *Function of mask* (Fmask) [16]. It was originally designed for Landsat imagery but later extended for Sentinel-2 data [15]. Here, we use the Fmask version as implemented in FORCE ([17]), which is able to separate clouds from bright surfaces exploiting parallax effects. In FORCE, the cloud masking is integrated into a processing workflow, which also includes coregistration [18], radiometric correction [19], resolution merging [20] and datacube generation [21]. The individual detectors of MSI-sensor have slightly different viewing directions alternating from forward view to backward view between adjacent detectors. The second code is the latest version of ATCOR (v 9.3.0), which contains a masking algorithm [22] as a necessary preprocessing part before starting the atmospheric correction. Masking in ATCOR 9.3.0 was improved relative to previous versions. The third is the scene classification of Sen2Cor (version 2.8.0). Sen2Cor is an atmospheric correction processor for Sentinel-2 (S2) data provided by the *European Space Agency* (ESA), which contains a preprocessing scene classification step preceding atmospheric correction [23]. Whereas the atmospheric correction module of Sen2Cor was developed in heritage of ATCOR, the scene classification is completely independent. Scene classification of Sen2Cor makes use of some external auxiliary data from Climate Change Initiative [24]. It is still to mention that Fmask uses a 300 m dilation buffer for cloud, and 60 m for cloud shadow, while ATCOR uses 100 m and 220 m, respectively, and Sen2Cor (version 2.8.0) uses no dilation buffers. Fmask applies also a 1 pixel buffer for snow. The reader is referred to the given references for a detailed description of the three methods and the different threshold values used, because it is outside the scope of this paper.

This paper is organized as follows: Section 2 presents an overview over the S2 scenes used for the exercise. Section 3 describes the approach to define the reference (“truth”) mask (validation procedure). Section 4 presents the classification results in terms of user’s, producer’s and overall accuracy [25], and Section 5 provides a discussion of the critical issues. The conclusion and possible further improvements are given at the end of the paper.

## 2. Methods (Processors) and Data

Twenty S2 scenes are processed with the three codes. A list of the investigated Sentinel-2 scenes is given in Table 2. The scenes were selected to cover all continents, different climates, seasons, weather conditions, and land cover classes (Figure 1). They represent flat and mountainous sites with cloud cover from 1% to 62% and include the presence of cumulus, thin and thick cirrus clouds and snow cover. Additionally, the scenes represent different land cover types such as desert, urban, cropland, grass, forest, wetlands, sand, coastal areas and glaciers. The range of solar zenith angles is from 18° to 62°. For the scene classification validation, all S2 bands with 10 m and 60 m are resampled to a common 20 m pixel size. All processors used *Digital Elevation Models* (DEMs) usually from SRTM (90 m) (downloaded from the USGS website (<https://earthexplorer.usgs.gov/>)) except for the scenes number 1, 6 and 16, which used Planet DEM [26].

The classification of ATCOR provides a map with 22 classes which is used in the subsequent atmospheric correction module [22]. For this investigation, a compact map with seven classes (clear, semitransparent cloud, cloud, cloud shadow, water, snow/ice, topographic shadow) is derived from the detailed map at 20 m spatial resolution. A potential shadow mask is defined as reference shadow and the cloud height of the cloud mask is iterated until the projected cloud mask for the given solar geometry matches the shadow mask. Topographic shadow is calculated with a ray tracing algorithm using the DEM and the solar geometry. The classes “cloud shadow” and “water” are often difficult to distinguish and in case of cloud shadow over water the class assignment is arbitrary. Therefore, misclassifications can happen, because only one label can be assigned

in this method. Semitransparent cloud can be thin cirrus or another cloud type of low optical thickness.

The SCL algorithm of Sen2Cor aims to detect clouds with their shadows and to generate a scene classification map. The latter raster map consists of 12 classes, including 2 classes for cloud probabilities (medium, and high), thin cirrus, cloud shadows, vegetated pixels, nonvegetated pixels, water, snow, dark feature pixels, unclassified, saturated or defective pixels and no data. This map is used internally in Sen2Cor in the atmospheric correction module to distinguish between cloudy pixels, clear land pixels and water pixels, and it does not constitute a land cover classification map in a strict sense [27]. The scene classification map is delivered at 60 m and 20 m spatial resolution, with associated *Quality Indicators* (QI) for cloud and snow probabilities. The QIs provide the probability measure (0–100%) that the Earth surface is obstructed either by clouds or by snow. Class dark area pixels can contain dark features like burned area, topographic shadows or cast shadows but also very dark water bodies and vegetation. Thin cirrus may also be other transparent cloud and the transition from medium to high probability cloud is impossible to validate. Pixels assigned to unclassified are mostly pixels with low probability of clouds or mixed pixels, which do not fit into any of the other classes.



**Figure 1.** Geographical distribution of 20 test sites selected for validation (orange squares).



**Table 2.** Sentinel-2 level L1C test scenes (SZA = Solar Zenith Angle). Information on scene cloud cover, climate, main surface cover, rural/urban.

Scene	Location	Date	Tile	SZA	Cloud Cover	Desert	Ice/Snow	Nonveg	Veg	Water	Mountains	Rural	Urban
1	Antarctic	2019/02/03	T21EVK	54.9°	28%		X	X		X			
2	Argentina, Buenos Aires	2018/08/27	T21HUC	51.5°	0%					X		X	X
3	Australia, Lake Lefroy	2018/08/19	T51JUF	51.5°	0%			X					
4	Bolivia, Puerto Siles	2018/09/06	T19LHF	30.6°	0%				X	X		X	
5	China, Dunhuang	2018/01/22	T46TFK	62.3°	24%	X	X					X	
6	Estonia, Tallin	2018/07/14	T35VLG	39.0°	2%					X		X	X
7	Germany, Berlin	2018/05/04	T33UUU	38.0°	1%			X	X	X		X	X
8	Italy, Etna	2017/03/09	T33UUU	45.1°	7%		X			X		X	X
9	Kazakhstan, Balkhash	2018/07/30	T43TFM	30.7°	7%	X				X		X	
10	Mexico, Cancun	2018/05/27	T16QDJ	18.4°	7%				X	X		X	
11	Morocco, Quarzazate	2018/08/30	T29RPQ	27.2°	2%	X		X	X	X			
12	Mosambique, Maputo	2018/07/13	T36JVS	54.4°	0%					X		X	X
13	Netherlands, Amsterdam	2018/09/13	T31UFU	49.7°	5%				X	X		X	X
14	Philippines, Manila	2018/03/19	T51PTS	27.4°	1%					X		X	X
15	Russia, Sachalin	2018/05/09	T54UVC	35.5°	0%		X	X		X			
16	Russia, Yakutsk	2017/08/08	T52VEP	45.9°	6%			X	X	X			
17	Spain, Barrax-1	2017/05/09	T30SWH	24.1°	18%				X		X	X	
18	Spain, Barrax-2	2017/05/19	T30SWH	22.0°	2%				X		X	X	
19	Switzerland, Davos	2019/04/17	T32TNS	37.7°	25%		X	X	X	X	X		
20	USA, Rimrock	2018/05/12	T11TMM	30.4°	1%		X			X		X	X

### 3. Validation Procedure

Validation of masks comprises verification of the mask classification accuracy to clarify uncertainties of masking products for their applications. Comparison of different mask classification algorithms requires first to map all the individual masking outputs to a common set of labels. Table 3 shows the seven classes used as a common set for Fmask, ATCOR and Sen2Cor.

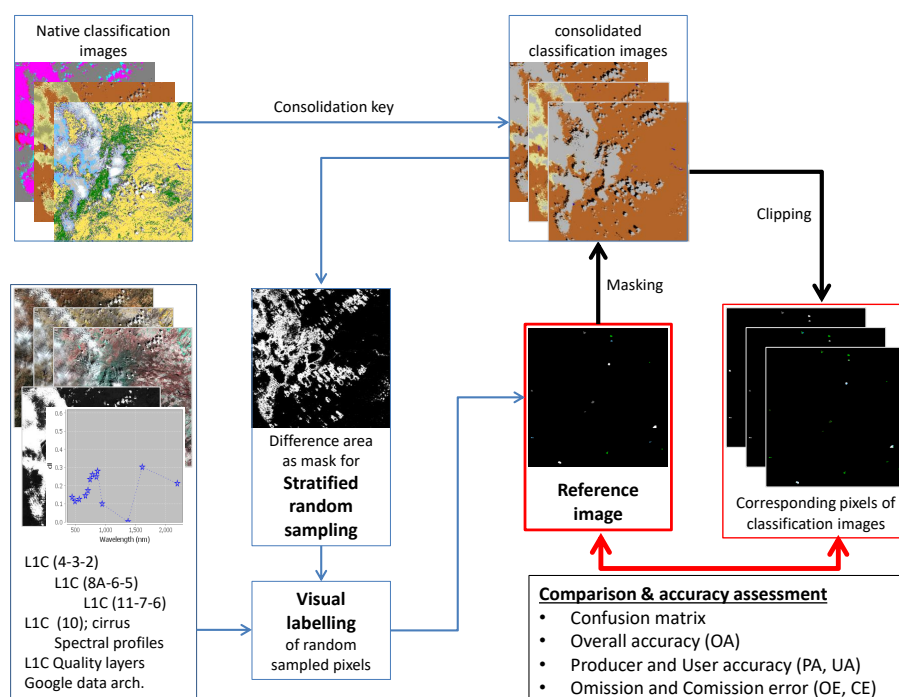
Semitransparent cloud is defined as optically thin cirrus cloud, thin lower altitude cloud, haze or smoke. To detect thin cirrus clouds, the reference mask generation use the TOA reflectance in the cirrus band 10, lying below 0.04 but above 0.01. The lower threshold is used to avoid classifying all pixels as semitransparent. The label cloud comprises optically thick (opaque) water cloud and also cirrus cloud with  $\rho$  (TOA, band 10)  $> 0.04$ .

The focus of the present paper is not only validation of the scene classification provided by the three processors but its comparison. This comparison is done by generating two reference maps which complement each other—one for the “difference area” and another for the “same area”. The difference area is the part of the classification images where the classification maps provided by the three processors disagree. Validation statistics over the difference area enable a relative comparison between processors pointing on strengths and weaknesses much sharper than interpreting statistics over an entire image, which are often fairly similar. The same area is the remaining part of the images where all three classifications give the same result. Combination of the validation statistics over the same area and disagreement area enables to assess the absolute classification performance of the processors. This requires that the ratio of labeled pixels in the difference area to the labeled pixels in the same area is the same as the ratio of the size of disagreement area to size of agreement area. The challenge for validation of SCL is generation of high quality reference maps which gives the “truth”. Generation of the reference maps for the performed comparison of Fmask, ATCOR and Sen2Cor outputs relies on visual inspection, supplemented by meteorological data, if available. The following procedure was repeated for each image of the validation dataset.

First, stratified random sampling [25] is applied to the difference mask between three processors to get the sample points for visual labeling. Stratification serves to get the sample size balanced between all classes present in the image, thus to guarantee statistical consistency and to avoid exclusion of spatially limited classes from the validation. Our aim is an amount of 1000 randomly selected samples per image with the minimum number of 50 samples for the smallest class (for reference please see following authors: [28–31]). Visual inspection by human interpreter results in labeling of either one pixel only or alternatively labeling a polygon drawn around an adjacent area of pixels of the same class to assign the correct class and create the reference (“truth”) map. All labeled pixels are used to create the reference classification image typically resulting in an average number of 5000 pixels per scene. Figure 2 presents an overview on the generation of the classification reference mask. It begins (left part) with selected L1C channel combinations (4-3-2; 8A-6-5; 10; spectral TOA reflectance profiles, etc.), continues with the consolidation, stratified random sampling and visual labeling to create the reference image. This image (right part) is masked and compared to the consolidated images to obtain the corresponding pixels of the classification images and perform the accuracy assessment.

**Table 3.** Consolidation of individual masking outputs to common mask labels. Reference mask and corresponding mask of selected codes.

Label	Masks	Definition for Reference	Fmask	ATCOR	Sen2Cor
1	Clear land	Clear pixels over land	Clear	Clear	Vegetation; not vegetated; unclassified
2	Semitransparent cloud	$0.01 < \text{TOA } \rho(1.38 \mu\text{m}) < 0.04$ ; also haze, smoke or any kind of cloud which transparency enables to recognize the background features	Cloud	Semitransparent cloud	Thin cirrus
3	Cloud	Cumulus cloud; thick clouds (also thin cirrus)	Cloud	Cloud	Cloud medium and high probability
4	Cloud shadow	Shadow thrown by the clouds over land	Cloud shadow	Shadow	Cloud shadow
5	Clear water	Clear pixels over water	Water	Water	Water
6	Clear snow/ice	Clear pixels over snow and ice	Snow/Ice	Snow/ice	Snow and ice
7	Topographic shadow	Self-shadow and/or cast-shadows	-	Topographic shadow	Dark feature
0	Background		background	Geocoded background	No data; saturated or defective



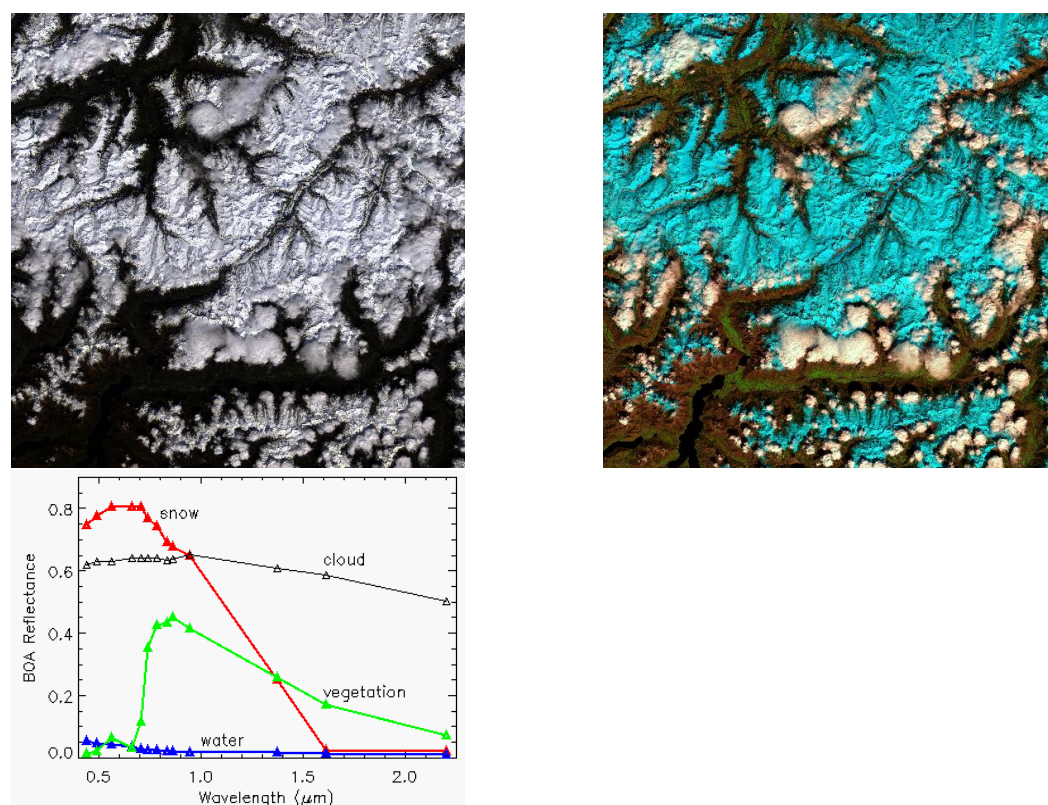
**Figure 2.** Schema for classification reference mask generation on example of Sen2Cor scene classification (SCL) product over Barrax test-site (Spain), acquired on 19 May 2017. This example represents various topography (flat and rough) and land-cover (vegetated, nonvegetated, water) as well as cloud cover dominated by the cumulus clouds. Process starting in left part with L1C channel combinations, continuing with the consolidation, stratified random sampling and visual labeling to create the reference image. Red circled image cubes: zoomed area.

Visual inspection by the expert human interpreter was supported by:

- Visual checks of the TOA true color image (bands 4, 3, 2), TOA near infrared false color (bands 8A, 6, 5), and TOA short-wave infrared false color (bands 12, 11, 8A).
- Check of L1C cirrus (band 10) concerning semitransparent cirrus regions.
- Check of BOA reflectance spectral profiles from Level-2 Sen2Cor products.
- Comparison with imagery archive from GoogleEarth<sup>TM</sup>.

The created reference classification map is finally compared to the consolidated classification maps from Fmask, ATCOR and Sen2Cor and a confusion matrix is obtained for each classification. Finally, classification accuracy statistics are computed from confusion matrices. After completing analysis for disagreement area, the same procedure is repeated for the same area to allow computation of absolute classification accuracy statistics of the three classifications.

Figure 3 shows an example of a true color (RGB = Red, Green, Blue = bands 4, 3, 1) composite of scene 19 (Davos) of Table 2, a false color composite using RGB (SWIR1, NIR, red) and some typical BOA reflectance spectra of snow, clouds, clear (vegetation) and water. Obviously, snow/ice and clouds cannot easily be discerned in the true color image. Therefore, the human interpreter also uses other band combinations, in this case with band 11 (SWIR1), where snow/ice (colored blue) is clearly recognized. In addition, BOA reflectance spectra are evaluated for a polygon if a class assignment is not obvious.



**Figure 3.** Left to right: true color (RGB = 665,560,443 nm) composite of scene ID 19, SWIR1 (RGB = 1600,860,660 nm) composite and example spectra.

The procedure applied for generation of our reference classification map is similar to the way used to create the references for the Hollstein and PixBox datasets [10]. The new point is that we split the validation into creating a reference for the difference area and another for the same area for comparison of classification tools. Please note that obtained reference maps are not perfect. The manual labeling includes some amount of subjectivity. Most of all visual interpretation and labeling of transparent clouds is challenging. Subjectivity of the method was tested with four people, creating a reference map for the same two products. The test revealed quite stable results with 5–6% differences in *overall accuracy* (OA) using the reference maps for computation of classification accuracy statistics. Another limitation of our classification reference maps comes from the stratified random sampling. The stratification between classes has to be oriented itself on one classification, which was Sen2Cor in our case. If the Sen2Cor classification fails, then the reference map becomes imbalanced. Even if this is not the case, then the reference maps are not perfectly balanced for the other classifications. The potential bias could be investigated by creating another stratified random sampling based on a different set, but such a sensitivity study is outside the scope of this paper.

Classification accuracy statistics is represented by three parameters calculated from the error confusion matrix [25,29,30]. If the number of classes is  $n$ , then the confusion matrix  $C$  is a  $n \times n$  matrix, and the *user's accuracy* of class  $i$  (percentage of the area mapped as class  $i$  that has reference class  $i$ ) is defined as

$$UA(i) = 100 C_{ii} / \sum_{j=1}^n C(i, j) \quad j = \text{column number} \quad (1)$$



The second parameter is the *producer's accuracy* of class  $i$  (percentage of the area of reference class  $i$  that is mapped as class  $i$ )

$$PA(i) = 100 C_{ii} / \sum_{j=1}^n C(j, i) \quad j = \text{row number} \quad (2)$$

The last is the *overall accuracy*:

$$OA = 100 \frac{\sum_{j=1}^n C(j, j)}{\sum_{i=1}^n \sum_{j=1}^n C(i, j)} \quad (3)$$

The OA can be calculated for the total area of an image, i.e., the absolute OA but also for the difference and same area of each scene and masking code.

Besides the OA, UA and PA measures, a detailed visual inspection supported the analysis of the confusion within and between classes per processor. Comparison was performed per processor and class over difference area, including recognition rates, misclassification rates of particular class as well as its confusion potential with other classes (the proportion of one mistaken by other class).

#### 4. Results

Validation results consist of confusion matrix with the number of correctly classified pixels in the validation set. Confusion matrix is the basis for computation of UA and PA and OA of classification. Table 4 provides results for difference area and shows a summary of the UA and PA per class, i.e., the average over all 20 scenes. Table 5 provides results for absolute validation of classifications comparable to results present in the literature and contains the OA per scene. Boldface numbers indicate the method with the best performance, but if the values differ less than about 1% then two methods are marked correspondingly.

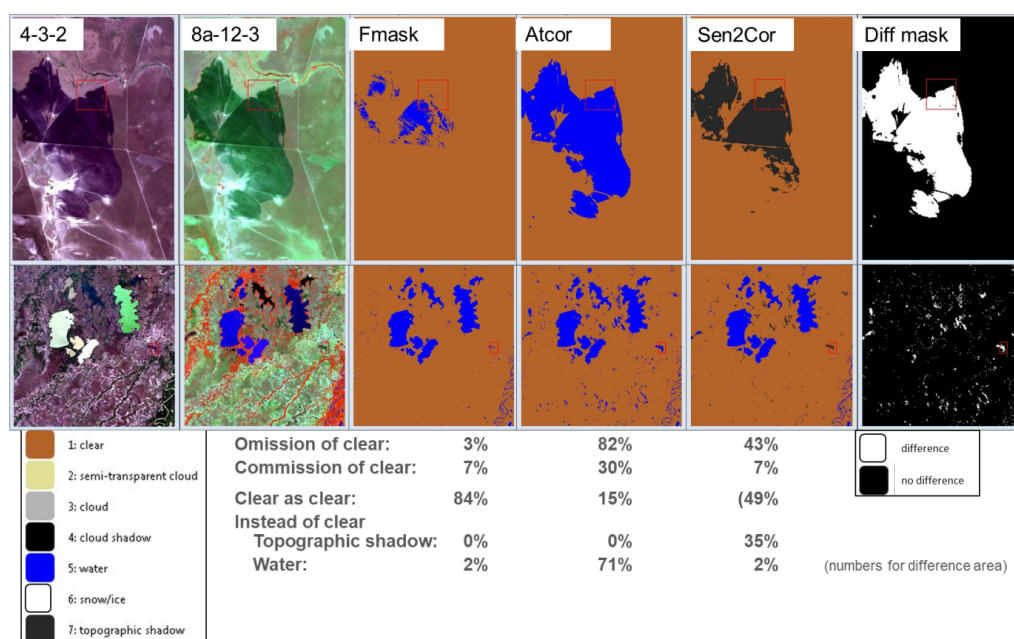
**Table 4.** Summary of classification accuracy (percent) for difference area (F = Fmask, A = ATCOR, S = Sen2Cor; bold face numbers indicate the best performances).

Class	UA (F)	UA (A)	UA (S)	PA (F)	PA (A)	PA (S)
clear	79.8	69.0	<b>80.8</b>	56.2	64.6	<b>75.5</b>
semitransp. cloud	<b>78.2</b>	36.4	67.1	1.8	<b>30.4</b>	<b>28.2</b>
cloud	13.4	<b>39.3</b>	34.6	<b>84.5</b>	62.7	65.7
cloud shadow	50.8	53.3	<b>82.3</b>	42.8	<b>49.1</b>	27.7
water	<b>94.1</b>	44.8	70.0	68.1	52.9	<b>80.3</b>
snow/ice	53.0	<b>58.9</b>	49.5	25.2	75.7	<b>85.7</b>
topographic shadows	<b>75.9</b>	43.0	14.6	2.2	4.1	<b>53.0</b>

**Table 5.** Summary of overall accuracy (percent) (F = Fmask, A = ATCOR, S = Sen2Cor).

Scene	Location	OA Difference Area			OA Same Area			OA Total Area		
		F	A	S	F	A	S	F	A	S
ID	Average (all scenes)	45	56	<b>62</b>	97	97	97	89	91	92
1	Antarctic	36	51	<b>56</b>	95	98	100	79	86	<b>88</b>
2	Argentina, Buenos Aires	<b>90</b>	59	59	98	98	98	<b>98</b>	96	96
3	Australia, Lake Lefroy	59	49	<b>67</b>	100	100	100	96	95	<b>97</b>
4	Bolivia, Puerto Siles	<b>90</b>	22	58	100	100	100	99	98	<b>99</b>
5	China, Dunhuang	56	40	<b>74</b>	97	97	100	73	64	<b>85</b>
6	Estonia, Tallin	40	71	<b>78</b>	98	98	99	90	94	<b>96</b>
7	Germany, Berlin	57	<b>76</b>	67	100	100	100	97	<b>98</b>	<b>98</b>
8	Italy, Etna	32	<b>70</b>	<b>71</b>	100	100	100	88	<b>95</b>	<b>95</b>
9	Kazakhstan, Balkhash	47	<b>75</b>	45	92	91	91	87	<b>90</b>	87
10	Mexico, Cancun	44	59	<b>66</b>	99	99	99	89	92	<b>93</b>
11	Morocco, Quarzazate	71	<b>86</b>	45	100	100	100	96	<b>98</b>	93
12	Mosambique, Maputo	<b>75</b>	35	45	85	85	85	<b>84</b>	83	83
13	Netherlands, Amsterdam	38	<b>63</b>	<b>64</b>	96	96	96	86	90	<b>91</b>
14	Phillipines, Manila	46	<b>69</b>	67	98	98	98	92	<b>95</b>	94
15	Russia, Sachalin	<b>85</b>	51	63	99	98	98	<b>98</b>	93	94
16	Russia, Yakutsk	<b>64</b>	49	53	97	97	97	<b>94</b>	92	93
17	Spain, Barrax-1	25	<b>74</b>	62	99	99	99	83	<b>93</b>	91
18	Spain, Barrax-2	64	<b>92</b>	90	100	100	100	97	<b>99</b>	<b>99</b>
19	Switzerland, Davos	16	42	<b>46</b>	86	97	97	54	72	<b>74</b>
20	USA, Rimrock	50	52	<b>63</b>	99	99	99	98	<b>99</b>	98

For space reasons, we cannot present detailed results for each scene. The example of scene 4 (Bolivia, Puerto Siles) in Figure 4 serves as an example to demonstrate the difference mask validation. The image contains no clouds but water with different color and sediment, bright soil and burned area. This image is the example with the smallest difference area rather than the one with the largest agreement between Fmask, ATCOR and Sen2Cor classifications. This is also underlined with high absolute OA over complete image of 99%, 98% and 99%. There is only a small difference between classifications in PA over complete image for class clear land with 100%, 98% and 99% representing what is visible in Figure 4—a different amount of burned area is classified as water. User accuracy of class water for the total image is 99% for all masking codes, hiding differences clearly to see in the figure. Statistics over difference area gives a much more detailed insight into classification performance. OA over difference area is 90%, 22% and 58% for Fmask, ATCOR and Sen2Cor. Differences in PA for class clear land are now more highlighted with values 97%, 18% and 57%. User accuracy of class water for the difference image now is different between Fmask and Sen2Cor with 74% resp. 80%. Whereas Fmask identifies 97% of clear pixels in the difference area as clear, ATCOR and Sen2Cor do it for less than 60% of pixels. ATCOR largely misclassifies burned area as water. The problem shown for Sen2Cor with misclassification of clear land as topographic shadow has its origin in the transformation of Sen2Cor classification outputs to the consolidated mask. Consolidated class topographic shadow corresponds to Sen2Cor class dark area, which can contain dark features like burned area, topographic shadows or cast shadows but also very dark water bodies and vegetation. A planned update of class definition of Sen2Cor class dark area to only topographic or cast shadow will solve this confusion.

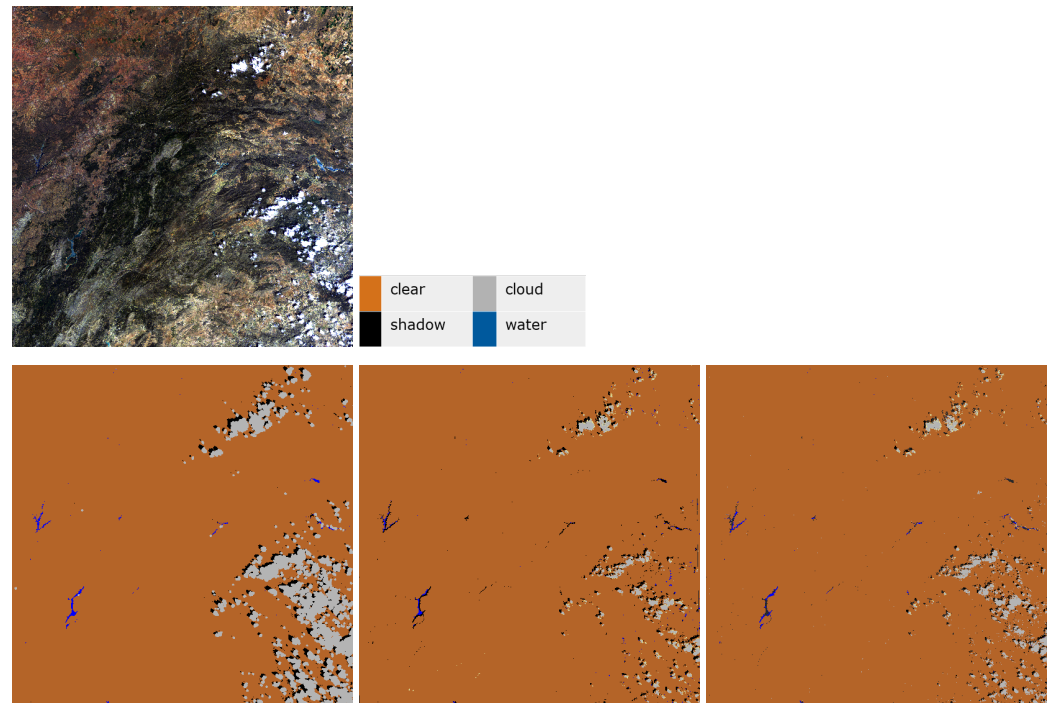


**Figure 4.** Difference area validation on example of scene 4 (Bolivia, Puerto Siles). **Bottom row:** Sentinel-2 Scene; **top row:** zoom of image showing a region with burned area; From left to right: Natural color composite of bands 2, 3, 4; false color composites of bands 8a, 12, 3 helpful for discrimination between dark classes, vegetation types and clouds; Classification map from Fmask; Classification output of ATCOR; Classification map from Sen2Cor; Difference area map.

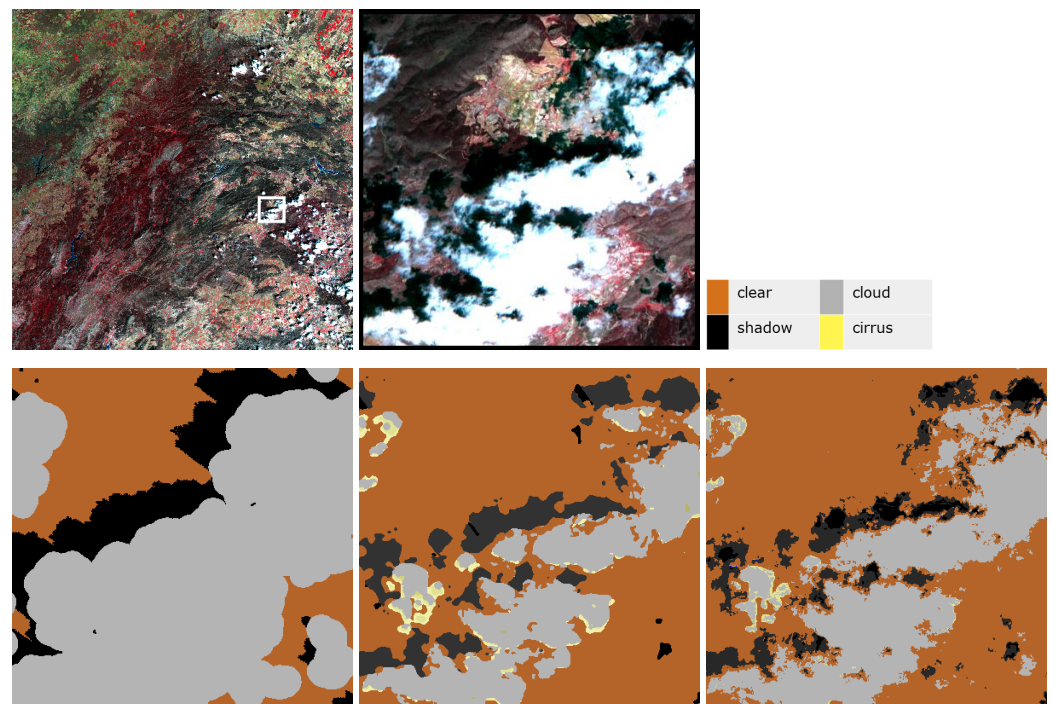
To furthermore compare the classification performance of Fmask, ATCOR and Sen2Cor, details are given for three selected cases: the best and worst case scenarios and an average case.

Figure 5 shows the best case (highest absolute overall accuracy) scenario of all analyzed scenes from Table 5. It is scene number 18 from Spain (Barrax) taken on 19 March 2017 with a zenith angle of 22.0° and a azimuth angle of 143.2°. In Figure 6 subset of scene ID

18 can be found. It nicely illustrates the differences between Fmask, ATCOR and Sen2Cor. The cloud percentage is overestimated in Fmask due to mask dilation, while ATCOR and Sen2Cor classifications are very similar and close to the reference.



**Figure 5.** Top row: true color (RGB = 665,560,443 nm) composite of scene ID 18 (Barrax-2). Bottom row (left to right): Fmask, ATCOR and Sen2Cor classification maps.



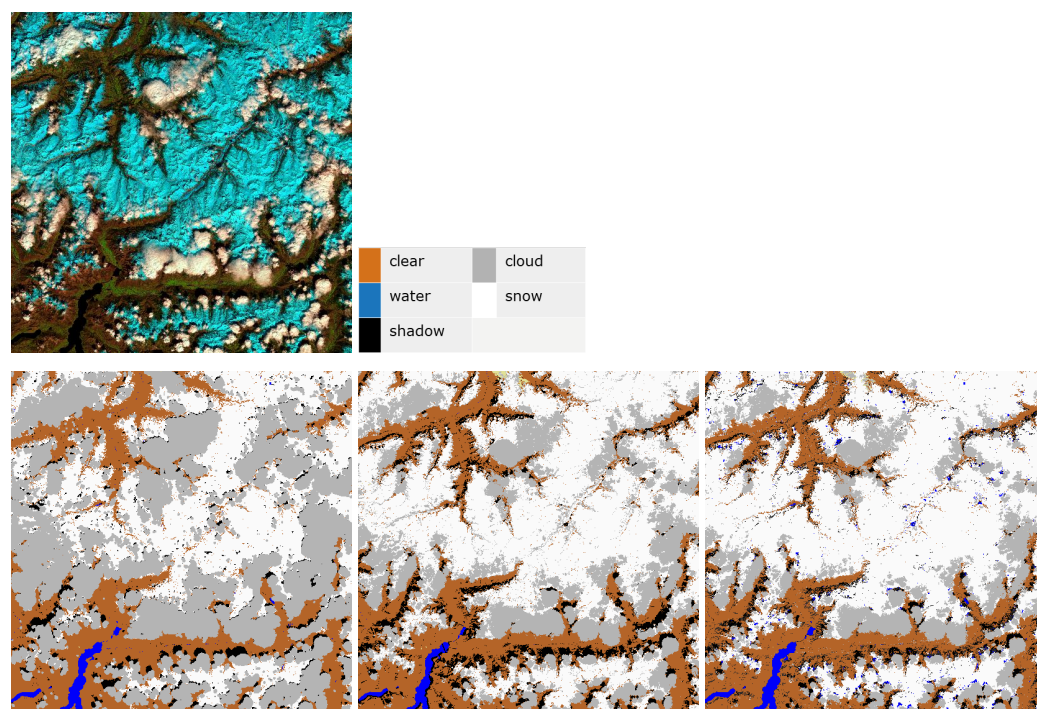
**Figure 6.** Top row (left to right): CIR (RGB = 865,665,560 nm) composite and CIR subset of scene ID 18 (Barrax-2). Bottom row (left to right): Fmask, ATCOR and Sen2Cor classification maps of the subset.

The overall worst case scenario (lowest absolute OA) of the 20 scenes analyzed is illustrated in Figure 7. This scene from Switzerland (Davos) was acquired on 4 April 2019 at a zenith and azimuth angle of  $37.7^\circ$  and  $158.5^\circ$ , respectively. This scene is difficult to classify correctly for all the processors due to the high reflectivity of the snow and complex topography. The snow is often misclassified as cloud. The lower overall accuracy in Fmask compared to ATCOR and Sen2Cor is again connected with the cloud dilation.

Figure 7 shows a subset of scene ID 19. As in the previous case (scene ID 18 from Spain) Fmask overestimates the percentage of cloud coverage at the expense of snow cover, which may or may not be problematic depending on the application. ATCOR and Sen2Cor show a more accurate cloud mask. An inspection of a zoom area (see Figure 8) reveals that Sen2Cor sometimes falsely classifies cloud shadows as water.

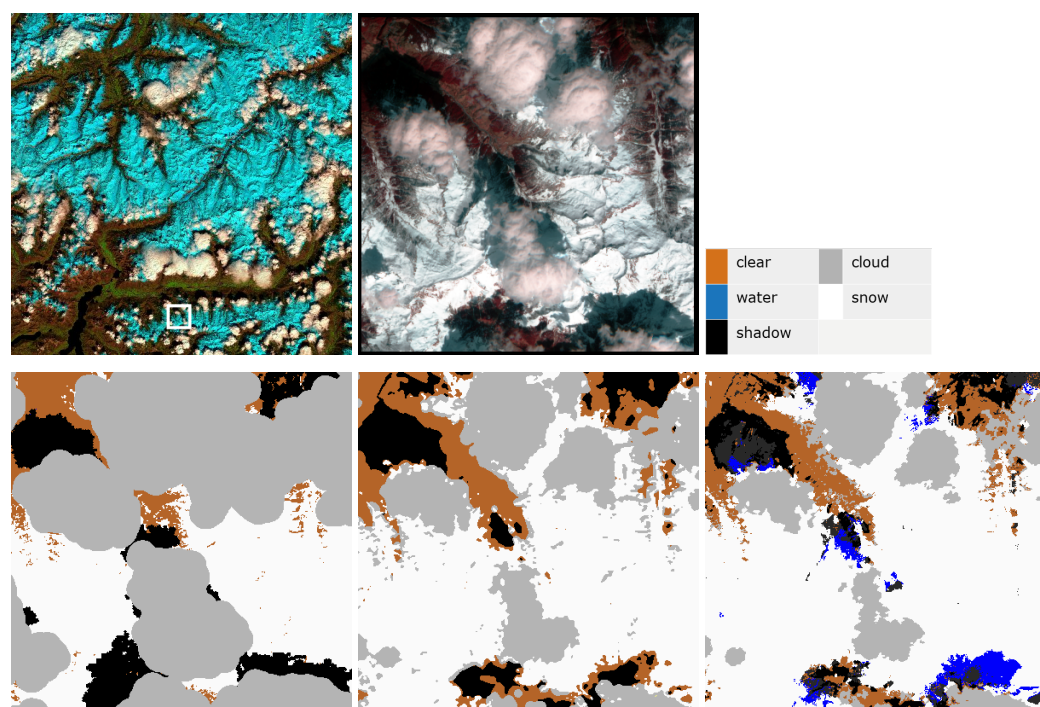
A scene showing an average case scenario (i.e., no complex topography, small percentage of cloud cover and bright objects) for all classification methods is the one from the USA (Rimrock). It was taken on 12 May 2018 at a zenith angle of  $30.4^\circ$  and an azimuth angle of  $153.5^\circ$ . Figure 9 shows the entire area of the scene with the three different classification maps, whereas Figure 10 only illustrates a subset of scene ID 20. Most of the scene is clear with some clouds and snow/ice in the southern part. Additionally, the river is accurately mapped by all processors.

The subset (Figure 10) demonstrates the difficulties Sen2Cor faces when distinguishing between urban areas or bright ground objects and clouds. ATCOR on the other hand misinterprets dark water for shadow. However, if both classes have about the same probability, then ATCOR's preference is shadow.

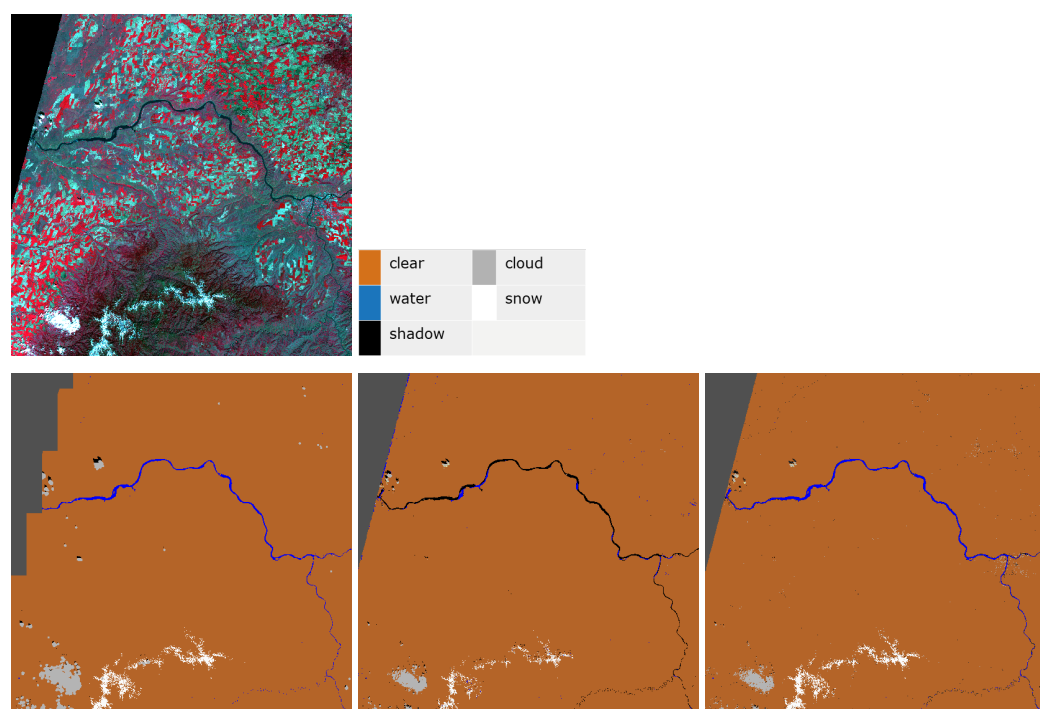


**Figure 7.** Top row: SWIR1/NIR/red composite of scene ID 19 (Davos). Bottom row (left to right): Fmask, ATCOR and Sen2Cor classification maps.

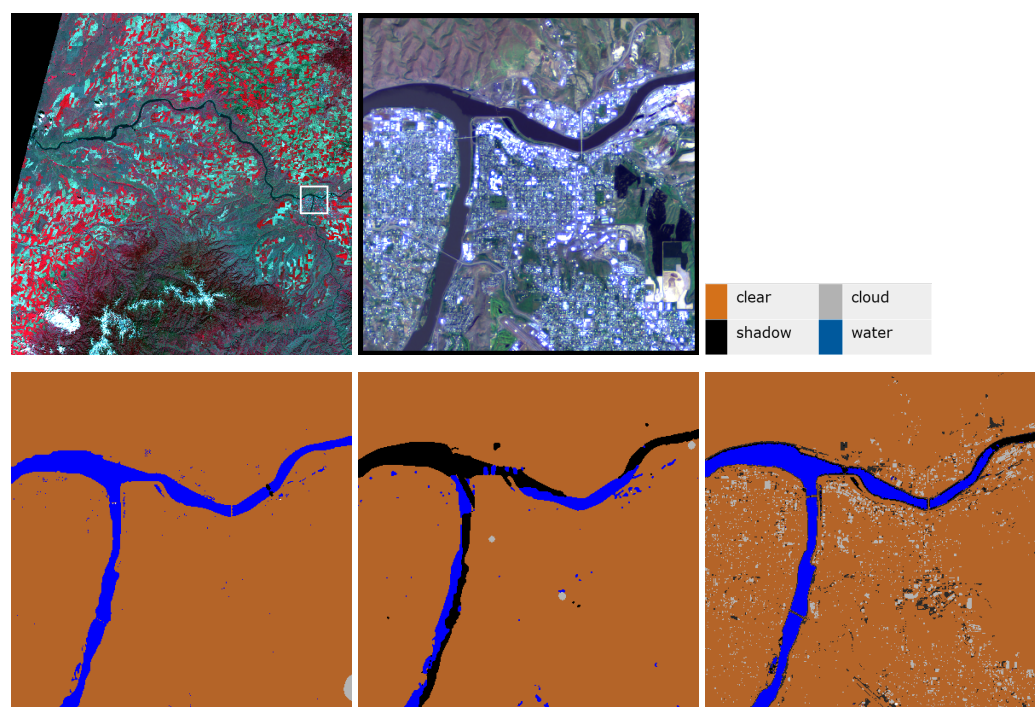




**Figure 8.** Top row (left to right): SWIR1/NIR/red composite and CIR (RGB = 865,665,560nm) subset of scene ID 19 (Davos). Bottom row (left to right): Fmask, ATCOR and Sen2Cor classification maps.



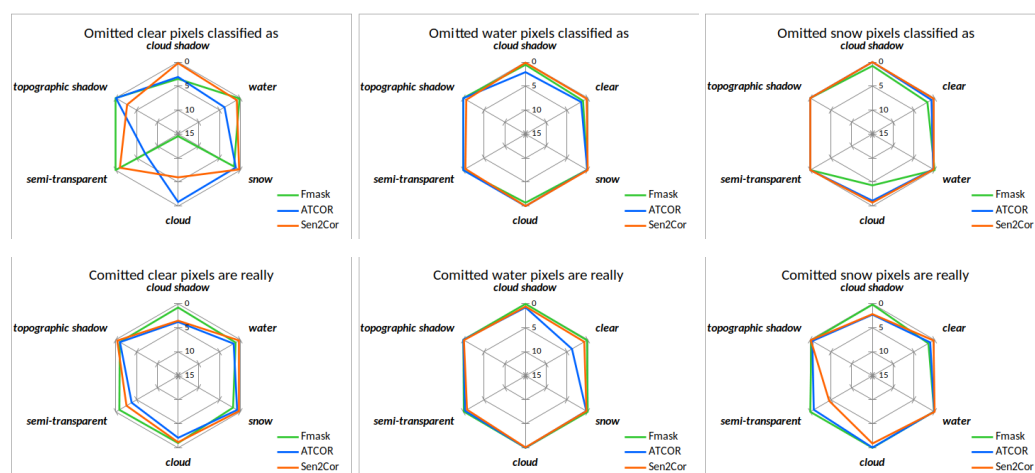
**Figure 9.** Top row: CIR (RGB = 865,665,560 nm) composite of scene ID 20 (USA Rimrock). Bottom row (left to right): Fmask, ATCOR and Sen2Cor classification maps.



**Figure 10.** Top row (left to right): CIR (RGB = 865,665,560 nm) composite and true color (RGB = 665,560,443 nm) subset of scene ID 20 (USA Rimrock). Bottom row (left to right): Fmask, ATCOR, and Sen2Cor classification maps.

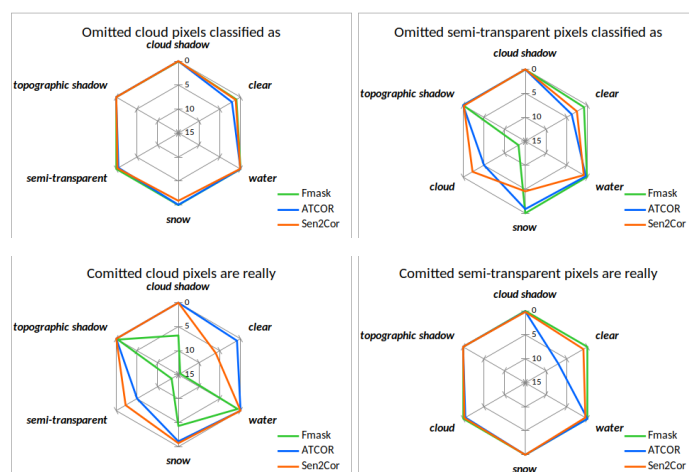
As can be deduced from Table 4 for the difference area, up to 75.5% of clear pixels were correctly classified by Sen2Cor, whereas Fmask and ATCOR recognize 56.2% and 64.6% correctly. The highest share of clear pixel misclassification was found by clouds for Fmask and Sen2Cor and semitransparent clouds for ATCOR. Semitransparent clouds were recognized up to 30.4% and 28.2% for ATCOR and Sen2Cor, respectively, while the omitted pixels were mainly distributed between classes clear and clouds by ATCOR and clear and snow by Sen2Cor. Fmask only classifies 1.8% of semitransparent cloud pixels correctly and mostly misclassifies the omitted pixels as clouds. Fmask performs best for the classification of cloud pixels (84.5%), while ATCOR and Sen2Cor have a recognition rate of 62.7% and 65.7%, respectively. The highest share of the cloud omission was found by class clear for Fmask and Sen2Cor and by class cloud shadows for ATCOR. Cloud shadows have low recognition rate (27.7%) and high confusion with class clear in the case of Sen2Cor. Fmask and ATCOR have lowest recognition for the class topographic shadows with a rate of 2.2% and 4.1%, respectively. Sen2Cor performs slightly better with 53.0%. Their omission is distributed mainly between classes clear and cloud shadows. The highest recognition rates (and lowest confusion to other classes) were found for clouds (84.5%) and water (68.1%), clear (64.6%) and snow (75.7%) and water (80.3%) and snow pixels (85.7%), for Fmask, ATCOR and Sen2Cor, respectively. Surprisingly, the proportion of snow pixels being mistaken toward clouds was low for ATCOR and Sen2Cor (12% and 8%, respectively), whereas Fmask misclassifies 47%, which is because of the cloud buffer, as well as the compilation of FORCE quality bits into the scene classification as employed in this study. In original FORCE output, multiple flags can be set for one pixel, i.e., the snow and cloud flags can both be set. During the reclassification process, clouds were given highest priority, thus snow detections were overruled by buffered cloud detections.

The confusion within and between classes can be additionally illustrated using the proportion of the individual class omissions for the difference area (Figures 11–13).



**Figure 11.** Omission and commission per Class for difference area for clear classes clear land, water and snow.

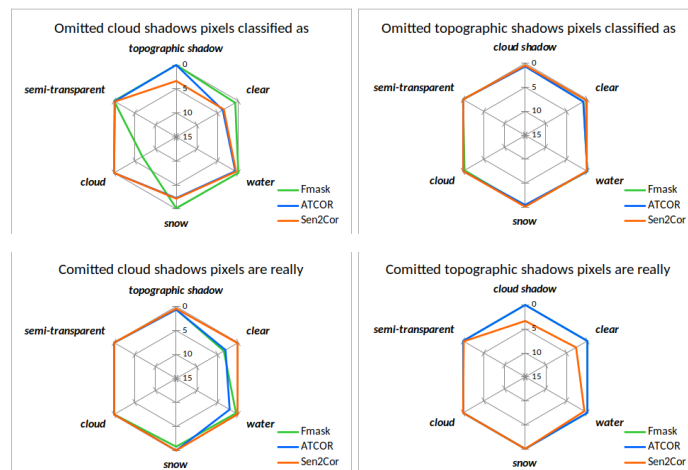
Figure 11 illustrates spider diagrams for the omission and commission of the classes clear land, water and snow representing valid, cloudless pixels. Fmask, ATCOR and Sen2Cor are represented by the colors green, blue and orange. Looking at the left upper plot of Figure 11, it can be noted that Fmask has a large omission of clear land towards clouds. This can be clearly attributed to cloud dilation. ATCOR has a omission of clear land to water, which is uncritical for pure cloud masking. Sen2Cor confuses most clear land pixels with topographic shadows due to the unfavorable class definition of class dark feature, which is mapped to topographic shadows. In the central lower plot of Figure 11 we see that ATCOR confuses most water pixels as clear land. All three masking codes show commission of water pixels towards the same direction of clear land but with different amounts. The commission of snow shows that Fmask and ATCOR classifies some clear as snow, which is uncritical for clear/cloud mask. Sen2Cor classifies some semitransparent clouds as snow.



**Figure 12.** Omission and commission per Class for difference area for cloud classes cloud and semitransparent cloud.

Figure 12 illustrates spider diagrams for the omission and commission of the classes cloud and semitransparent cloud for difference area. The upper left image shows that ATCOR and Sen2Cor have omission of cloud pixels towards the class clear. The commission of cloud is on the other hand different for all three masking codes. Fmask shows the largest commission of cloud pixels towards clear, semitransparent cloud and cloud shadow. Sen2Cor classifies some clear and semitransparent pixels as cloud and ATCOR shows a slight commission of semitransparent pixels towards cloud. For the class semitransparent

cloud, the largest omission comes from Fmask, which confuses most semitransparent clouds as cloud. This perfectly corresponds to commission of cloudy pixels towards semitransparent clouds. ATCOR and Sen2Cor show a commission of semitransparent pixels towards the class clear.



**Figure 13.** Omission and commission per Class for difference area for shadow classes cloud shadows and topographic shadows.

Figure 13 illustrates spider diagrams for the omission and commission for difference area of the shadow classes cloud shadows and topographic shadows. From the left upper image of figure 13 it can be noted that Fmask has the largest omission of cloud shadow towards the class cloud. ATCOR and Sen2Cor confuse cloud shadows mostly with clear pixels. All three masking codes show a similar direction of commission of cloud shadows towards clear pixels. Except for the class definition problem of Sen2Cor for topographic shadows, the upper and lower right images show good agreement between the processors and almost perfect performance. Sen2Cor shows a large commission of topographic shadow pixels towards the class clear, water and cloud shadow due to its definition of dark pixels.

## 5. Discussion

Since the reference and classified maps are based on the same dataset, i.e., a perfect match of geometry and acquisition time, the main uncertainty of the reference map classification is the use of a human interpreter [29]. Experiences with similar experiments using several human analysts report an average interpretation variability of ~5–7% [16,32] for cloud masks. In order to reduce the influence of the interpreter, a reference polygon should have homogeneous BOA reflectance properties per class, i.e., heterogeneous areas with mixed pixels are excluded [30]. The area homogeneity can be checked visually per band and it also shows if pixel spectra of a polygon have a large dispersion, e.g., for cloud border regions or snow areas below semitransparent cloud. Although the variability within a polygon should be small, large differences can exist between different polygons of the same class, e.g., in the case of different cloud types or fresh and polluted snow.

Table 4 presents the class-specific *user's accuracy* (UA) and *producer's accuracy* (PA) for the three methods averaged over the 20 scenes valid for difference area. High PA values (>80%) are only achieved for the classes cloud (Fmask) and snow/ice (Sen2Cor) indicating how difficult a classification is for all other classes. The low values for semitransparent cloud are most likely caused by the interpreter and his visual assessment, which does not agree with the selected physical criterion ( $0.01 < \rho(\text{TOA}, 1.38 \mu\text{m}) < 0.04$ ) of the three methods. Another known classification problem concerns the distinction of water and cloud shadow if no external maps are included. Both classes can be spectrally very similar.



Additionally, there can be cloud shadow over water, but since a pixel can only belong to one class in our evaluation, the setting of the preference rule adds another uncertainty.

Nevertheless, a comparison with S2 classification results obtained by the standard Fmask [15] (applied to seven scenes) demonstrates that, in our investigation, all three methods yield better overall accuracies than presented in reference [15] (Figure 6). This is even more remarkable because our approach uses six classes instead of four, and an increase of the number of classes usually tends to decrease the overall accuracy. One has to consider that the spatial resolution of Sentinel-2 data is 20 m, while it is 30 m for the Landsat data of reference [15]. A better classification agreement might, at least partly, be achieved by the enhanced Sentinel-2 resolution. However, while a higher spatial resolution can help to achieve a better classification, this is mainly related to mixed pixels, and in our study heterogeneous areas with mixed pixels are excluded.

Table 5 allows a selection of the best method depending on location and cloud cover:

- Fmask can best be applied for scenes in moderate climate, excluding arid and desert regions as well as areas with a large snow/ice coverage.
- ATCOR can best be applied for urban (bright surfaces), arid and desert scenes.
- Sen2Cor can best be applied for rural scenes in moderate climate and also in scenes with snow and cloud.

Again, a reminder is needed: the Fmask results shown here pertain to the Fmask parallax version [17] not the available standard version [15]. Furthermore, Sen2Cor uses an additional external ESACCI-LC data package, which improves the classification accuracy over water, urban and bare areas and enables a better handling of false detection of snow pixels [33]. Therefore, Sen2Cor benefits from a certain advantage compared to Fmask and ATCOR. During this investigation we also found out that the performance of Fmask (parallax version) can be improved if the current cloud buffer size of 300 m is reduced to 100 m. In the meantime, the size of the cloud buffer has become a user-defined parameter. The performance of Sen2Cor (version 2.8.0) can be slightly improved with an additional cloud buffer of 100 m (instead of no buffer), whereas an additional 100 m cloud buffer is almost of no influence on the ATCOR performance.

To sum up, we can say that the overall accuracy is very high for all three masking codes and nearly the same (89%, 91% and 92% for Fmask, ATCOR and Sen2Cor, respectively) and the balanced OA (OA for same area) is equal (97%). ATCOR finds most valid pixels, has the highest PA and lowest UA for valid pixels. Sen2Cor finds less valid pixels due to its class definition of dark area. Fmask finds least valid pixels due to dilation of cloud masks, thus not a randomly occurring commission. In contrast, Fmask has the lowest cloud omission and clear commission at the expense of higher cloud commission and clear omission. Depending on application, losing a higher rate of cloud-adjacent pixels may be far less severe than missing cloud pixels.

## 6. Conclusions

The performance of three classification methods (Fmask, parallax version), ATCOR and Sen2Cor was evaluated on a set of 20 Sentinel-2 scenes covering all continents, different climates, seasons and environments. The reference maps with seven classes (clear, semitransparent cloud, cloud, cloud shadow, water, snow/ice and topographic shadows) were created by an experienced human interpreter. The average overall accuracy for the absolute area is 89%, 91%, and 92% for Fmask, ATCOR, and Sen2Cor, respectively. High values of producer's accuracy of the difference area (>80%) were achieved for cloud and snow/ice, and lower values for the other classes typically range between 30% and 70%. This study can serve as a guide to learn more about possible pitfalls and achieve more accurate algorithms. Future improvements for the classification algorithms could involve texture measures and convolutional neural networks.



**Author Contributions:** V.Z.: validation, writing of article; M.M.-K.: concept, methodology, validation; J.L.: software, writing; D.F.: software, writing; R.R.: software, writing; B.P. concept, methodology, writing. All authors have read and agreed to the published version of the manuscript.

**Funding:** Part of this research was performed as part of the Copernicus Sentinel-2 Mission Performance Center activities, which are managed by ESA. This research received no other external funding.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
2. Immitzer, M.; Vuolo, F.; Atzberger, C. First experience with Sentinel-2 data for crop and tree species classifications in Central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
3. Clevers, J.G.; Gitelson, A.A. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. *Int. J. Appl. Earth Obs. Geoinform.* **2013**, *23*, 344–351. [CrossRef]
4. Hagolle, O.; Huc, M.; Pascual, D.V.; Dedieu, G. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. *Remote Sens. Environ.* **2010**, *114*, 1747–1755. [CrossRef]
5. Muller-Wilm, U.; Louis, J.; Richter, R.; Gascon, F.; Niezette, M. Sentinel-2 level 2A prototype processor: Architecture, algorithms and first results. In Proceedings of the ESA Living Planet Symposium, Edinburgh, UK, 9–13 September 2013.
6. Yan, L.; Roy, D.P.; Zhang, H.; Li, J.; Huang, H. An automated approach for sub-pixel registration of Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multi Spectral Instrument (MSI) imagery. *Remote Sens.* **2016**, *8*, 520. [CrossRef]
7. Reddy, B.S.; Chatterji, B.N. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Trans. Image Process.* **1996**, *5*, 1266–1271. [CrossRef] [PubMed]
8. Baraldi, A.; Tiede, D. AutoCloud+, a “Universal” Physical and Statistical Model-Based 2D Spatial Topology-Preserving Software for Cloud/Cloud-Shadow Detection in Multi-Sensor Single-Date Earth Observation Multi-Spectral Imagery—Part 1: Systematic ESA EO Level 2 Product Generation at the Ground Segment as Broad Context. *Int. J. Geo-Inf.* **2018**, *7*, 457. [CrossRef]
9. Defourny, P.; Bontemps, S.; Bellemans, N.; Cara, C.; Dedieu, G.; Guzzonato, E.; Hagolle, O.; Inglada, J.; Nicola, L.; Savinaud, M.; et al. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. *Remote Sens. Environ.* **2019**, *221*, 551–568. [CrossRef]
10. Hollstein, A.; Segl, K.; Guanter, L.; Brell, M.; Enesco, M. Ready-to-Use Methods for the Detection of Clouds, Cirrus, Snow, Shadow, Water and Clear Sky Pixels in Sentinel-2 MSI Images. *Remote Sens.* **2016**, *8*, 666. [CrossRef]
11. Paul, F.; Winsvold, S.H.; Käab, A.; Nagler, T.; Schwaizer, G. Glacier remote sensing using Sentinel-2. Part II: Mapping glacier extents and surface facies, and comparison to Landsat 8. *Remote Sens.* **2016**, *8*, 575. [CrossRef]
12. Du, Y.; Zhang, Y.; Ling, F.; Wang, Q.; Li, W.; Li, X. Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band. *Remote Sens.* **2016**, *8*, 354. [CrossRef]
13. Earth.esa.int. 2020. ACIX II—CMIX 2Nd WS. Available online: <https://earth.esa.int/web/sppa/meetings-workshops/hosted-and-co-sponsored-meetings/acix-ii-cmix-2nd-ws> (accessed on 5 June 2020).
14. Baetens, L.; Desjardins, C.; Hagolle, O. Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure. *Remote Sens.* **2019**, *11*, 433. [CrossRef]
15. Zhu, Z.; Wang, S.; Woodcock, C.E. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sens. Environ.* **2015**, *159*, 269–277. [CrossRef]
16. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94. [CrossRef]
17. Frantz, D.; Hass, E.; Uhl, A.; Stoffels, J.; Hill, J. Improvement of the Fmask algorithm for Sentinel-2 images: separating clouds from bright surfaces based on parallax effects. *Remote Sens. Environ.* **2018**, *215*, 471–481. [CrossRef]
18. Rufin, P.; Frantz, D.; Yan, L.; Hostert, P. Operational Coregistration of the Sentinel-2A/B Image Archive Using Multitemporal Landsat Spectral Averages. *IEEE Geosci. Remote. Sens. Lett.* **2020**, 1–5. [CrossRef]
19. Frantz, D.; Röder, A.; Stellmes, M.; Hill, J. An Operational Radiometric Landsat Preprocessing Framework for Large-Area Time Series Applications. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3928–3943. [CrossRef]
20. Frantz, D.; Stellmes, M.; Röder, A.; Udelhoven, T.; Mader, S.; Hill, J. Improving the Spatial Resolution of Land Surface Phenology by Fusing Medium- and Coarse-Resolution Inputs. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4153–4164. [CrossRef]
21. Frantz, D. FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. *Remote Sens.* **2019**, *11*, 1124, 10.3390/rs11091124. [CrossRef]

22. Richter, R.; Schlöpfer, D. *ATCOR Theoretical Background Document*; DLR Report DLR-IB 564-03/2019; German Aerospace Center (DLR): Wessling, Germany, 2019. Available online: <https://www.rese-apps.com/software/atcor/manual-papers.html> (accessed on 15 January 2020).
23. Louis, J. TO BE UPDATED Sentinel 2 MSI—Level 2A Product Definition. Issue 4.4. 2016-08-12. Available online: <https://sentinel.esa.int/documents/247904/1848117/Sentinel-2-Level-2A-Product-Definition-Document.pdf> (accessed on 15 January 2020).
24. Hollmann, R.; Merchant, C.J.; Saunders, R.; Downy, C.; Buchwitz, M.; Cazenave, A.; Chuvieco, E.; Defourny, P.; de Leeuw, G.; Holzer-Popp, T.; et al. The ESA Climate Change Initiative. Satellite Data Records for Essential Climate Variables. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 1541–1552. [[CrossRef](#)]
25. Congalton, R.G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* **1991**, *37*, 35–46. [[CrossRef](#)]
26. Planet DEM. Available online: <https://www.planetobserver.com/products/planetdem/planetdem-30/> (accessed on 25 September 2018).
27. Gascon, M.; Zijlema, W.; Vert, C.; White, M.; Nieuwenhuijsen, M. Outdoor blue spaces, human health and well-being: A systematic review of quantitative studies. *Int. J. Hyg. Environ. Health* **2017**, *220*, 1207–1221. [[CrossRef](#)] [[PubMed](#)]
28. Giles M. Foody Sample size determination for image classification accuracy assessment and comparison. *Int. J. Remote Sens.* **2009**, *30*, 5273–5291. . [[CrossRef](#)]
29. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [[CrossRef](#)]
30. Stehman, S.V. Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* **2009**, *30*, 5243–5272. [[CrossRef](#)]
31. Wagner, J. ; Stehman, S. Optimizing sample size allocation to strata for estimating area and map accuracy. *Remote Sens. Environ.* **2015**, *168*, 126–133. [[CrossRef](#)]
32. Oreopoulos, L.; Wilson, M.J.; Varnai, T. Implementation on Landsat Data of a Simple Cloud-Mask Algorithm Developed for MODIS Land Bands. *IEEE GRSL* **2009**, *30*, 5243–5272. [[CrossRef](#)]
33. S2 MPC—Sen2Cor Configuration and User Manual. Ref.S2-PDGS-MPC-L2A-SUM-V2.8. Issue 2. 2019-02-05. Available online: <http://step.esa.int/thirdparties/sen2cor/2.8.0/docs/S2-PDGS-MPC-L2A-SUM-V2.8.pdf2> (accessed on 14 February 2020).